

Article

End-to-End Deep Learning Pipeline for Multi-Sensor Aircraft Engine Vibration Fault Diagnosis

Yijun Xie ^{1,2,3} , Jiaxian Sun ^{1,2,*}, Chunyan Hu ^{1,2,3} , Haoran Pan ^{1,2}, Chenchen Wang ^{1,2,3}  and Junqiang Zhu ^{1,2,3}

¹ Key Laboratory of Light-Duty Gas-Turbine, Institute of Engineering Thermophysics, Chinese Academy of Sciences, Beijing 100190, China; xieyijun@iet.cn (Y.X.); huchunyan@iet.cn (C.H.); panhaoran@iet.cn (H.P.); wangchenchen@iet.cn (C.W.); zhujq@iet.cn (J.Z.)

² National Key Laboratory of Science and Technology on Advanced Light-Duty Gas-Turbine, Beijing 100190, China

³ School of Space Exploration, University of Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: sunjiaxian@iet.cn

Abstract

Aero-engine safety and prognostics and health management (PHM) rely on robust vibration-based fault diagnosis. However, many deep learning studies on rotating machinery are evaluated under random train–test splits that mix hardware instances and may obscure the domain shift faced in deployment. This paper presents a protocol-driven end-to-end baseline for multi-sensor aero-engine-relevant vibration diagnosis on the HIT inter-shaft bearing benchmark. Six synchronous vibration channels are segmented into fixed-length windows, standardized using source-domain statistics, and classified by a compact 1D CNN backbone with and without squeeze-and-excitation (SE) channel attention. A deeper ResNet1D baseline is further introduced to examine whether increasing backbone capacity improves cross-bearing generalization under the same source-only training protocol. We compare random segment-level splits with bearing-level cross-splits that hold out entire bearings as unseen target domains, and we report deployment-oriented indicators including balanced accuracy, false-alarm rate (FAR), and miss rate over five random seeds. Under random splits, the compact CNN baseline reaches near-ceiling test accuracy, confirming that the benchmark is readily separable under in-domain interpolation. In contrast, cross-bearing evaluation reveals severe degradation: in the representative split, the baseline CNN accuracy collapses to approximately 15% with near-zero normal-class recall, while ResNet1D improves fault sensitivity but still retains a high FAR above 88%. Additional cross-bearing permutations further show that this degradation is not attributable to a single unfavorable source–target split. These findings indicate that, under the tested source-only backbones and protocols, distribution mismatch is a dominant bottleneck for deployment-ready cross-bearing diagnosis. The results establish a reproducible baseline for protocol-driven evaluation in aero-engine PHM and motivate future work on domain adaptation, domain generalization, calibration, and sequential decision logic.



Academic Editor: Sujin Bureerat

Received: 18 May 2026

Revised: 15 June 2026

Accepted: 22 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: aero-engine vibration; inter-shaft bearing; deep learning; domain shift; fault diagnosis

1. Introduction

Aero-engines operate under high rotational speed, high thermal load, and strong aerodynamic excitation, where vibration is both an unavoidable dynamic response and a key observable for condition monitoring and safety assurance [1–3]. In industrial practice,

vibration-based monitoring is widely adopted because it is non-intrusive, responsive to mechanical anomalies, and compatible with on-line trending and alarm systems [4]. However, aero-engine vibration is typically multi-source and non-stationary, and the measured signals often reflect coupled effects from shafts, bearings, casings, and flow-path excitations rather than a single isolated component [3,5]. This coupling complicates reliable early warning, especially for incipient faults whose signatures can be weak compared with operating-condition variability and measurement noise [6].

Among vibration-related anomalies, rotor–stator rubbing and bearing-related faults are especially relevant because they can escalate quickly and may manifest as intermittent, nonlinear, and transient features [7,8]. Classical diagnosis pipelines therefore rely on signal processing and hand-crafted indicators to highlight impulsive or modulated components [6,9]. These methods are well-established and interpretable, yet their performance can be sensitive to parameter choices, operating conditions, and sensor placement; moreover, feature engineering and classifier design are often decoupled, which limits end-to-end optimization [9,10]. As a result, there is growing interest in data-driven learning systems that directly map raw or minimally processed vibration to health states, with reduced reliance on manual feature construction [11,12].

Deep learning, particularly convolutional architectures, has shown strong capability in extracting hierarchical representations from time-series signals for fault diagnosis [12,13]. Batch normalization and modern stochastic optimization further improve training stability and convergence for such models [14,15]. In addition, channel-attention mechanisms such as squeeze-and-excitation (SE) blocks can adaptively recalibrate feature channels and have demonstrated effectiveness in improving representation quality in deep networks [16]. However, the evaluation protocol used to assess such models is often as important as the model architecture itself.

A central concern is that many rotating-machinery fault diagnosis studies rely on random segment-level splits. When vibration recordings are divided into overlapping windows, random splitting can place segments from the same physical bearing, operating condition, or adjacent time interval into both training and test sets. Such a protocol mainly evaluates in-domain interpolation and can produce overly optimistic accuracy, especially when bearing-specific or run-specific signatures are shared between the two subsets [17]. This issue is not merely statistical. For aero-engine PHM, the deployed model is expected to operate on new bearings, new assemblies, or new engines whose transfer paths and nuisance factors were not observed during training. Therefore, a model that performs well under a random split may still fail under an unseen-bearing deployment condition.

The broader machine-learning literature has long recognized this issue as dataset shift or domain mismatch [18,19]. Domain-adversarial learning and deep transfer learning have also been widely studied as possible remedies [20–22]. Nevertheless, for aero-engine-relevant multi-sensor vibration diagnosis, there remains a need for clear source-only baselines and rigorous evaluation protocols that explicitly separate in-domain interpolation from cross-bearing generalization. Without such baselines, it is difficult to determine whether high reported accuracy reflects deployable diagnostic evidence or merely favorable data splitting.

For aero-engine PHM, several benchmarking datasets have accelerated algorithmic development, including turbofan degradation simulation sets such as C-MAPSS and N-CMAPSS [23,24]. Meanwhile, rotating machinery diagnosis research frequently uses bearing datasets for method validation [25]. However, these widely used datasets may not reflect aero-engine multi-sensor vibration coupling and inter-shaft transmission characteristics that are central to whole-engine vibration monitoring [3,5]. Hou et al. released an inter-shaft bearing benchmarking dataset acquired from an aero-engine test rig with vibra-

tion signals measured on rotors and casings, providing a rare opportunity to study fault diagnosis under aero-engine-relevant transfer paths and cross-bearing generalization [26].

In this paper, we position the study as a protocol-driven and deployment-oriented baseline rather than as a new domain-adaptation algorithm. The primary objective is to quantify how strongly the apparent performance of end-to-end vibration diagnosis models depends on the evaluation protocol. We use the HIT inter-shaft bearing benchmark to compare conventional random segment-level splits with stricter bearing-level cross-splits that hold out entire bearings as unseen target domains. This design allows us to expose the gap between in-domain performance and cross-bearing deployability, while avoiding target-domain statistic leakage during preprocessing.

To keep the study reproducible and interpretable, we implement a compact 1D CNN as a transparent baseline, an SE-enhanced variant to examine the effect of channel-wise recalibration, and a ResNet1D baseline to examine whether a deeper residual backbone can alleviate the observed degradation under the same source-only training protocol.

The main contributions of this work are as follows:

- We formulate aero-engine-relevant multi-sensor vibration diagnosis as a cross-bearing generalization problem and explicitly separate random segment-level evaluation from unseen-bearing evaluation.
- We construct a reproducible source-only pipeline from six-channel time-domain segmentation and leakage-free normalization to 1D convolutional classification, providing controlled baselines with HitCNN1D, HitCNN1D_SE, and ResNet1D.
- We show that random segment-level splits can yield near-ceiling performance while cross-bearing evaluation exposes severe false-alarm risk and unstable generalization. Additional cross-bearing permutations further indicate that this degradation is not an artifact of a single source–target split.
- We provide deployment-oriented analysis using balanced accuracy, false-alarm rate, miss rate, and class-wise recall, and we identify domain adaptation, domain generalization, calibration, and sequential decision logic as necessary directions for future deployable PHM systems.

2. Aero-Engine Dataset and Problem Formulation

2.1. HIT Inter-Shaft Bearing Benchmark

The experimental study is conducted on the HIT inter-shaft bearing vibration benchmark proposed by Hou et al. [26]. To the best of our knowledge, this benchmark is among the first publicly described system-level aero-engine datasets for inter-shaft bearing fault diagnosis. It simultaneously records rotor and casing vibration responses under inter-shaft bearing fault conditions. Commonly available bench-top bearing rigs usually isolate a single bearing–housing subsystem. In contrast, the HIT benchmark was acquired on an aero-engine-based test platform with coupled rotor–casing dynamics and realistic transmission paths. Therefore, it contains strong broadband interference, structural resonances, and operating-condition variability that are typically absent in component-isolated laboratory datasets [26].

This distinction is important for deployable diagnosis. The dataset paper reports that bearing characteristic frequencies are not directly identifiable in the conventional spectrum for the HIT benchmark, whereas they are comparatively clearer in widely used datasets such as the CWRU bearing data and the accelerated degradation datasets typified by XJTU-SY [25–27]. This property makes the benchmark suitable for evaluating whole-engine-relevant monitoring methods, because fault signatures are masked by system-level dynamics, sensor mounting variability, and complex propagation channels [10,12].

2.1.1. Sensors and Measurement Channels

The benchmark provides six synchronous vibration channels acquired at representative locations of the rotor–casing system. Two displacement sensors measure the horizontal and vertical vibration displacement responses of the low-pressure rotor, and four acceleration sensors measure the casing vibration acceleration responses near the inter-shaft bearing support [26]. This heterogeneous sensing configuration is consistent with practical aero-engine monitoring, where rotor displacement and casing acceleration channels capture complementary dynamic responses. It also provides a natural setting for evaluating multi-sensor fusion architectures [12,26].

2.1.2. Engineering Monitoring Context

For each bearing instance, the raw acquisition is a long time series sampled at 25 kHz. Hou et al. report that the released benchmark retains fixed-length short sequences of 20,480 samples after segmentation and invalid-data removal [26]. In the public data organization, channels 1–2 correspond to rotor displacement responses and channels 3–6 correspond to casing acceleration responses. Each sample is associated with operating-speed metadata and a fault label in the benchmark definition [26].

In this paper, the model input is a six-channel time-domain window rather than hand-crafted indicators such as RMS, FFT peaks, or envelope-spectrum features. This end-to-end input design keeps the inference interface close to an on-line monitoring pipeline. It also avoids masking distribution mismatch through feature engineering, so that cross-bearing performance degradation can be attributed more directly to domain shift in the raw multi-sensor measurements. The model output is a per-window fault posterior, which supports two practical PHM actions: window-level alarming for rapid operational awareness and optional logging of high-confidence abnormal windows for subsequent maintenance analysis. While passenger comfort is secondary to safety for inter-shaft bearing faults, early identification of abnormal vibration signatures is still important because it can prevent fault escalation and structural resonance [1,2,10].

2.1.3. Fault States

The dataset contains three inter-shaft bearing states: healthy bearing, inner-race defect, and outer-race defect [26]. Although these faults are physically localized at the bearing, their observed vibration responses are system-level mixtures transmitted through the rotor–casing structure. The model must therefore infer the latent health state from complex coupled measurements rather than from clean component-isolated signatures. This makes the benchmark a useful proxy for whole-engine vibration diagnosis and a suitable testbed for cross-bearing generalization analysis [10,26].

2.2. Multi-Sensor Vibration Segmentation

To construct supervised learning examples suitable for deep sequence models, the raw multi-sensor recordings are segmented into fixed-length windows. Following the benchmark setting, we segment each channel into windows of $T = 20,480$ samples with 50% overlap [26]. The overlap increases the effective sample count and reduces boundary artifacts, but it also introduces strong temporal correlation between adjacent windows. Consequently, how the dataset is split becomes a first-order methodological choice, for segment-level random splits can unintentionally place highly correlated windows into both training and test sets, inflating apparent performance and underestimating deployment risk [17]. With a 25 kHz sampling rate, each window spans approximately 0.82 s, and the 50% overlap implies an effective decision update period of about 0.41 s, which is consistent with low-latency on-line alarming in vibration monitoring systems [1,26].

Each segmented sample is represented as a multi-channel time series:

$$\mathbf{X} \in \mathbb{R}^{C \times T}, \quad (1)$$

where $C = 6$ is the number of sensor channels and $T = 20,480$ is the number of time points per segment. In the processed dataset, each segment constitutes a short multi-sensor snapshot of the coupled rotor–casing vibration response and is directly fed to end-to-end temporal convolutions without handcrafted feature extraction.

After segmentation, we obtain 2412 labeled segments in the three-class setting, with a moderately imbalanced distribution: 954 normal segments, 1008 inner-race fault segments, and 450 outer-race fault segments. For binary experiments, we either merge inner- and outer-race defects into a single fault class for Experiment 0, or restrict the task to normal vs. inner-race segments for Experiment 2a to isolate a clean cross-bearing generalization setting.

2.3. Fault Classes and Diagnosis Tasks

We formulate two diagnostic tasks that correspond to different layers of an aero-engine vibration health monitoring pipeline.

2.3.1. Binary Fault Detection

Given a multi-sensor segment $\mathbf{X}$, the binary detection task determines whether the system is normal or faulty. This task aligns with on-line alarming, where stable false-alarm behavior and high fault sensitivity are critical for operational monitoring [10,12]. Experiment 0 studies this task under a random segment-level split by merging inner-race and outer-race defects into a single fault class. Experiment 2a studies the normal-vs.-inner-race version of the same task under cross-bearing evaluation.

2.3.2. Fault-Type Discrimination

The three-class task further discriminates among normal state, inner-race defect, and outer-race defect. This task is relevant when the monitoring system is expected not only to detect abnormality but also to provide fault-type information for maintenance decision making. The HIT benchmark specifically targets inter-shaft bearing faults, but the methodological objective of this study is broader: to examine whether representations learned from multi-sensor vibration remain reliable under complex signal propagation

and unseen-bearing domain shift. Experiment 1 evaluates three-class discrimination under a random segment-level split, while Experiment 2b extends the cross-bearing setting to a larger label space as discussed in Section 4.5.

2.4. Random and Cross-Bearing Split Evaluation Protocols

The evaluation protocol is not a minor implementation detail. It directly determines whether reported performance reflects deployable generalization or only in-domain interpolation. We therefore distinguish two protocols throughout this study.

2.4.1. Random Segment-Level Splits

In many rotating machinery studies, segmented windows are randomly divided into training, validation, and test subsets. This protocol is convenient, but it can be overly optimistic when sliding-window segmentation is used. Segments from the same bearing and operating condition may appear in both training and test sets, and adjacent overlapping windows may even be separated across these subsets. The learned model can therefore exploit bearing-specific or run-specific cues, yielding near-ceiling accuracy that primarily reflects in-domain interpolation rather than true unseen-bearing generalization [12,17].

2.4.2. Cross-Bearing Splits

A cross-bearing protocol is adopted to approximate the deployment regime involving unseen bearings or new engine assemblies. In this protocol, training and validation are restricted to source-domain bearings, while one or more target-domain bearings are held out exclusively for final testing. This design induces an explicit distribution shift between the source bearings and target bearings and is therefore more aligned with whole-engine monitoring practice [26].

Formally, this setting corresponds to the dataset-shift problem,

$$p_{\text{train}}(\mathbf{X}, y) \neq p_{\text{test}}(\mathbf{X}, y), \quad (2)$$

which is a well-known failure mode for data-driven monitoring systems [18]. It is also closely related to transfer learning and domain adaptation, where the central objective is to learn representations that remain stable across domains with different operating conditions, hardware instances, or nuisance factors [19,20,22]. In Section 4, we show that the dramatic performance gap between random splits and cross-bearing splits quantitatively exposes the severity of domain shift in system-level aero-engine vibration data.

2.4.3. Backbone-Agnostic Evaluation

The split design is model-agnostic. In addition to the lightweight HitCNN1D backbone and its SE-enhanced variant, we include a deeper residual 1D convolutional baseline, ResNet1D, under the same cross-bearing protocol. This setting separates two factors that are often confounded in vibration diagnosis studies: limited backbone capacity and distribution mismatch across hardware instances. By holding the evaluation protocol fixed while strengthening the backbone, we can examine whether cross-bearing degradation is reduced by a deeper source-only model or whether the dominant bottleneck remains source–target domain shift.

3. Proposed End-to-End Deep Learning Pipeline

3.1. Overall System Architecture

The proposed pipeline is designed to reflect an on-line vibration monitoring workflow for aero-engine condition monitoring while remaining computationally lightweight. In practical PHM systems, diagnostic models must process streaming sensor measurements, update health-state estimates with low latency, and avoid excessive feature-engineering overhead. Therefore, the pipeline is organized as an end-to-end mapping from synchronized multi-sensor vibration windows to window-level diagnostic decisions.

The architecture consists of four stages. First, six synchronous vibration channels are acquired from rotor displacement sensors and casing acceleration sensors and buffered into sliding windows [26]. Second, the buffered signals are segmented into fixed-length windows as described in Section 2.2. Channel-wise normalization is then applied using statistics estimated only from the training and validation data, so that target-domain information is not leaked into preprocessing. Third, each normalized window is processed by a temporal 1D deep backbone. We consider three source-only backbones under the same evaluation framework: a lightweight 1D CNN, its SE-enhanced variant, and a ResNet1D baseline. Fourth, global pooling and a linear classifier produce class logits and window-level diagnostic decisions.

3.1.1. Engineering Monitoring Interface and Input–Output Definition

From an engineering viewpoint, the HIT benchmark provides six synchronous vibration channels that mimic an aero-engine rotor–casing monitoring configuration [26].

Two channels correspond to rotor-related displacement measurements, and four channels correspond to casing-mounted acceleration measurements near the inter-shaft bearing support. The acquisition is sampled at 25,000 Hz in the dataset description [26]. Because the released files do not consistently report physical units for all channels, and because displacement and acceleration channels inherently differ in unit and dynamic range, we treat the released sequences as calibrated sensor outputs and standardize them through the deterministic normalization described in Section 3.2.

The model input is a six-channel time-domain window $\mathbf{X} \in \mathbb{R}^{6 \times 20480}$. The model output is a per-window posterior probability over health states, followed by a diagnostic decision after thresholding or class selection. This output supports two practical PHM actions: rapid window-level alarming for operational awareness and optional logging of high-confidence abnormal windows for subsequent maintenance analysis. The present study focuses on condition monitoring and fault diagnosis rather than remaining-life prediction.

3.1.2. Time-Domain Windows and Cross-Bearing Evaluation

The input is kept in the time domain rather than being converted into handcrafted RMS, FFT, or envelope-spectrum features. With a sampling frequency of 25,000 Hz and a window length of $T = 20,480$, each input window spans 0.8192 s of vibration data. With 50% overlap, the monitoring system can update the window-level probability and alarm decision every 0.4096 s under steady streaming. This update period provides a concrete latency reference for PHM-oriented on-line condition monitoring.

Using time-domain windows preserves phase-consistent transient and modulation patterns that may be attenuated by fixed feature pipelines. More importantly for this study, it enables a direct stress test of source–target distribution mismatch under structured splits. In deployment, the model must face new bearings or new assemblies whose transfer paths, sensor responses, and nuisance factors can differ from those observed during training, even when the nominal fault type is the same [19,20]. Therefore, cross-bearing evaluation is essential for assessing whether a model trained on source bearings can transfer to unseen target bearings.

3.1.3. Pipeline Summary

The key design principle is deployable end-to-end learning. Once segmentation and normalization are fixed, all remaining steps from raw multi-sensor windows to diagnostic decisions are learned jointly. This enables a controlled comparison between random segment-level evaluation and cross-bearing evaluation.

Figure 1 illustrates the proposed pipeline. The detailed architecture of the baseline HitCNN1D backbone is described in Section 3.3, and the ResNet1D configuration is summarized in Table 1.

Table 1. Detailed architecture of the ResNet1D baseline used in this study. The implementation uses a base channel width of 32.

Stage	Layer Type	Kernel	Stride	Output Channels
Input	Six-channel vibration window	–	–	6
Stem	Conv1D + BN + ReLU + MaxPool1D	7, 3	2, 2	32
Stage 1	$2 \times \text{BasicBlock1D}$	3	1	32
Stage 2	$2 \times \text{BasicBlock1D}$ with shortcut downsampling	3	2	64
Stage 3	$2 \times \text{BasicBlock1D}$ with shortcut downsampling	3	2	128
Stage 4	$2 \times \text{BasicBlock1D}$ with shortcut downsampling	3	2	256
Output	Global average pooling + fully connected layer	–	–	256

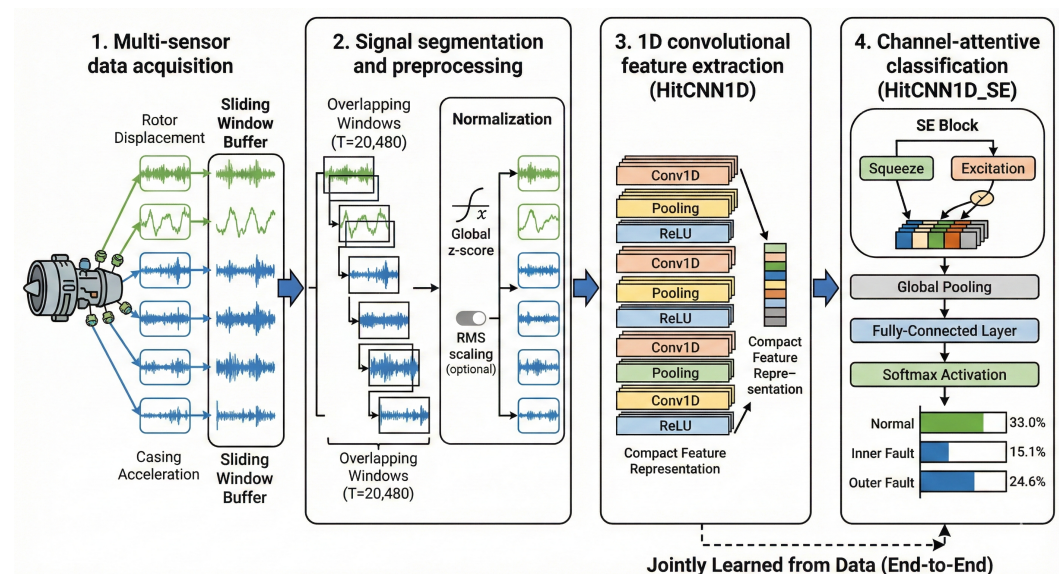


Figure 1. Proposed end-to-end vibration fault diagnosis system architecture diagram.

3.2. Signal Preprocessing and Normalization

Let $\mathbf{X} \in \mathbb{R}^{C \times T}$ denote a segmented multi-sensor window, as shown in Equation (1). We firstly perform channel-wise z-score normalization:

$$\tilde{\mathbf{x}}_{c,t} = \frac{\mathbf{x}_{c,t} - \mu_c}{\sigma_c}, \quad (3)$$

where μ_c and σ_c are estimated from the training+validation set for channel c . This choice is essential under cross-bearing evaluation, as using test-bearing statistics would implicitly leak target-domain information and distort the interpretation of domain shift [17,18]. Conceptually, global z-score normalization reduces sensor gain differences and stabilizes optimization, while preserving relative amplitude contrasts that may be diagnostically meaningful [12].

We further evaluate a per-segment RMS scaling:

$$\hat{\mathbf{x}}_{c,t} = \frac{\tilde{\mathbf{x}}_{c,t}}{\sqrt{\frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{x}}_{c,t}^2 + \epsilon}}, \quad (4)$$

which enforces unit energy per segment. The motivation is to reduce amplitude variability caused by operating speed differences or sensor mounting, thereby potentially improving cross-bearing transfer [18,19]. However, aggressive energy normalization may also suppress physically meaningful amplitude modulations associated with faults—particularly when the backbone lacks mechanisms to re-emphasize informative channels, and hence its net effect must be empirically validated as shown in Section 4.6.

3.3. Baseline 1D CNN Backbone

The baseline backbone HitCNN1D is a compact four-stage temporal CNN that balances representation capacity and computational efficiency. It uses strided convolutions and max pooling to progressively reduce temporal resolution while increasing the feature dimension [13,28]. Concretely, it comprises:

- Conv1: 16 filters, kernel size 5, stride 2, followed by BatchNorm and max pooling;
- Conv2: 32 filters, kernel size 5, stride 2, followed by BatchNorm and max pooling;
- Conv3: 64 filters, kernel size 3, stride 2, followed by BatchNorm and max pooling;
- Conv4: 128 filters, kernel size 3, stride 2, followed by BatchNorm and max pooling.

An adaptive average pooling aggregates the temporal dimension, producing a 128-dimensional representation that is mapped to logits by a linear classifier.

The role of HitCNN1D in this paper is diagnostic rather than purely competitive: it provides a transparent baseline to separate two factors often conflated in the literature: (i) in-domain performance under random splits and (ii) robustness under cross-bearing domain shift. As shown in Section 4, near-perfect scores under random splits do not translate to unseen-bearing settings, suggesting that source–target distribution shift is a major bottleneck under the tested source-only protocols.

3.4. ResNet1D Baseline with Residual Learning

To explicitly test whether increasing backbone capacity can mitigate cross-bearing domain shift under a source-only training protocol, we implement a ResNet-18-style 1D convolutional baseline [29]. The model follows the residual-learning principle of the original ResNet architecture, but all two-dimensional convolutional operations are replaced by one-dimensional temporal convolutions for multi-channel vibration sequences.

The ResNet1D implementation begins with a stem layer consisting of a 7×1 temporal convolution with stride 2, batch normalization, ReLU activation, and max pooling. It is followed by four residual stages, each containing two BasicBlocks. Each BasicBlock contains two 3×1 convolutional layers with batch normalization and ReLU activation. When the temporal resolution or channel dimension changes between stages, the shortcut branch uses a 1×1 convolution with the corresponding stride for downsampling. In our implementation, the base channel width is 32, so the four residual stages have 32, 64, 128, and 256 channels, respectively. Global average pooling is applied at the end to produce a 256-dimensional feature vector, which is mapped to the class logits by a fully connected layer.

Compared with the lightweight HitCNN1D backbone, the ResNet1D baseline has greater depth, residual connections, and a larger final representation dimension. Including this stronger backbone allows us to examine whether the cross-bearing degradation is primarily caused by insufficient representation capacity or by source–target distribution mismatch. We emphasize that this comparison is not intended to exhaust all possible deep architectures. Rather, it provides a controlled source-only baseline for evaluating whether a commonly used deeper residual backbone can reduce the observed cross-bearing degradation under the same data split, preprocessing, and model-selection protocol.

3.5. SE-CNN with Channel Attention

To examine whether channel-wise feature recalibration can improve robustness under heterogeneous sensing and cross-bearing domain shift, we augment HitCNN1D with squeeze-and-excitation (SE) modules and denote the resulting model as HitCNN1D_SE [16]. The backbone topology is kept identical to HitCNN1D, except that one SE block is inserted into each convolutional stage. Specifically, the placement is

$$\text{Conv1D} \rightarrow \text{BatchNorm} \rightarrow \text{ReLU} \rightarrow \text{SE block} \rightarrow \text{MaxPool1D}.$$

Thus, the SE block is applied after nonlinear feature extraction and before temporal downsampling. This placement allows the model to recalibrate channel responses before max pooling discards part of the temporal information.

For a feature map $\mathbf{F} \in \mathbb{R}^{K \times L}$ with K channels and temporal length L , the squeeze operation first applies global average pooling along the temporal dimension:

$$z_k = \frac{1}{L} \sum_{l=1}^L F_{k,l}, \quad k = 1, \dots, K. \quad (5)$$

The excitation operation then uses a two-layer bottleneck multilayer perceptron to generate channel-wise gates:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})), \quad (6)$$

where $\delta(\cdot)$ denotes the ReLU activation and $\sigma(\cdot)$ denotes the sigmoid function. The reduction ratio is set to 16 in all SE blocks. The recalibrated feature map is obtained by

$$\tilde{F}_{k,l} = s_k F_{k,l}. \quad (7)$$

In the context of multi-sensor vibration learning, SE attention provides a lightweight inductive bias for adaptively reweighting intermediate feature channels. This is relevant because different rotor and casing sensors can contain different mixtures of fault-related evidence and bearing-specific nuisance responses. However, SE attention is not treated as a domain-adaptation mechanism in this work. As shown in Section 4, its benefit under cross-bearing evaluation is limited and split-dependent, and it does not remove the false-alarm risk caused by source–target mismatch.

3.6. Training Configuration and Implementation Details

All models are implemented in PyTorch [30]. We optimize standard unweighted cross-entropy loss using Adam [15] and employ batch normalization to stabilize deep optimization [14]. Unless otherwise stated, the mini-batch size is 32 and the initial learning rate is 1×10^{-4} . Model selection is based on the validation loss: the checkpoint with the lowest validation loss is retained and evaluated on the held-out test domain. The test set is not used for early stopping, normalization-statistic estimation, or hyperparameter tuning.

For the source-only experiments reported in this paper, we do not use data augmentation, class-weighted loss, oversampling, undersampling, synthetic resampling, domain-adversarial learning, or target-domain calibration. This design choice keeps the experimental objective clear. The purpose is not to propose a final domain-robust diagnosis algorithm, but to establish a reproducible baseline and quantify how strongly the evaluation protocol affects the apparent deployability of end-to-end vibration diagnosis models.

The same preprocessing and source–target split definitions are used for HitCNN1D, HitCNN1D_SE, and ResNet1D. For cross-bearing experiments, normalization statistics are estimated only from the source-domain training and validation data and then applied to the target-domain test data. This prevents target-domain statistic leakage. Under this controlled setup, the comparison between the lightweight CNN-family models and the deeper ResNet1D baseline isolates the effect of backbone capacity under the present source-only training protocol.

4. Experimental Results

4.1. Common Experimental Settings

All experiments are implemented in PyTorch and trained with Adam optimization and cross-entropy loss. The mini-batch size is set to 32 and the initial learning rate is set to 1×10^{-4} . Model selection follows an early-stopping strategy based on the validation loss, and the checkpoint with the lowest validation loss is retained for test evaluation. Unless otherwise stated, the test set is never used for model selection, normalization-statistic estimation, or hyperparameter adjustment.

The experimental protocols are summarized in Table 2. For random-split experiments, all segmented windows are randomly divided at the segment level into training, validation, and test subsets with a stratified 64%/16%/20% ratio. This setting allows windows from the same physical bearing to appear in both training and testing, and therefore represents an in-domain interpolation protocol rather than a deployment-oriented generalization test. For

cross-bearing experiments, the split is performed at the bearing level. The source-domain bearings are used only for training and validation, while the target-domain bearings are held out for final testing.

Table 2. Summary of the experimental protocols and data splits. In the binary cross-bearing task, data1 and data2 correspond to normal bearings, while data3 and data4 correspond to inner-race fault bearings. data5 corresponds to the outer-race fault bearing.

Experiment	Task	Split Type	Training/Validation Data	Test Data
Experiment 0	Binary detection: normal vs. fault	Random segment-level split	All bearings; stratified train/validation split, 1543/386 segments	Randomly held-out test set, 483 segments
Experiment 1	Three-class diagnosis: normal vs. inner vs. outer	Random segment-level split	All bearings; stratified train/validation split, 1543/386 segments	Randomly held-out test set, 483 segments
Experiment 2	Binary detection: normal vs. inner	Cross-bearing split	Source domain: data1 normal + data3 inner; 806/202 train/validation segments	Target domain: data2 normal + data4 inner; 954 segments
Experiment 2B	Extended label-space cross-bearing evaluation	Cross-bearing split	Source domain: data1 normal + data3 inner + data5 outer; 1166/292 train/validation segments	Target domain: data2 normal + data4 inner; 954 segments
Permutation study	Binary detection: normal vs. inner	Four cross-bearing permutations	Four source-domain pairs; each split with a stratified 80%/20% train/validation ratio	Each target-domain pair is held out for testing; target size is 954 or 1008 segments depending on the permutation

For cross-bearing experiments, channel-wise z-score normalization is fitted using source-domain training and validation data only. The same source-domain statistics are then applied to the held-out target domain. No target-domain statistics are used during preprocessing. This design avoids test-domain information leakage and keeps the evaluation aligned with deployment, where the target bearing is not available during model development.

The class distribution is moderately imbalanced in the three-class setting because the outer-race class contains fewer segments than the normal and inner-race classes. To avoid masking this imbalance with overall accuracy alone, we report macro F1-score and class-wise recall. For all experiments reported in this paper, we use standard unweighted cross-entropy loss and do not apply class-weighted loss, oversampling, undersampling, or synthetic resampling. This choice keeps the study focused on the source-only evaluation protocol and avoids introducing additional balancing mechanisms that could confound the interpretation of cross-bearing domain shift.

For the cross-bearing binary setting, we evaluate accuracy, macro F1-score, and class-wise recall. In particular, normal-class recall is highlighted because it directly reflects the false-alarm tendency, which is a critical engineering concern for on-line monitoring.

In addition to overall accuracy and macro F1-score, we report three deployment-oriented indicators: balanced accuracy, false-alarm rate (FAR), and miss rate. For binary classification, these quantities are computed as

$$\begin{aligned} \text{Balanced Acc.} &= \frac{1}{2}(\text{Recall}_{\text{Normal}} + \text{Recall}_{\text{Fault}}), \\ \text{FAR} &= 1 - \text{Recall}_{\text{Normal}}, \quad \text{Miss} = 1 - \text{Recall}_{\text{Fault}}. \end{aligned} \quad (8)$$

These indicators are more aligned with on-line alerting than accuracy alone because they separate nuisance alarms from undetected faults.

4.2. Experiment 0: Binary Random Split

The first experiment aims to verify whether the proposed end-to-end pipeline can recover the simple decision boundaries in the HIT dataset when the training and test samples are drawn from the same distribution. To this end, all faulty samples containing inner and outer-race damage are merged into a single fault class, leading to a binary problem (normal vs. fault) under a random segment-level split across all bearings.

As shown in Table 3, under this setting, the baseline HitCNN1D already achieves essentially perfect performance, with test accuracy reaching 100% and both precision and recall being identical for the two classes. The corresponding confusion matrix is strictly diagonal, indicating that the model makes no observed misclassification on the held-out test set. This behavior shows that, when explicit cross-bearing shift is not imposed, the broadband multi-sensor signatures in the HIT dataset are highly separable by a relatively shallow 1D CNN.

Table 3. Experiment 0—random binary classification (normal vs. fault).

Class	Precision	Recall	F1-Score	Support
Normal	1.0000	1.0000	1.0000	191
Fault	1.0000	1.0000	1.0000	292
Overall accuracy	1.0000 (483/483)			

From an engineering perspective, Experiment 0 serves two important purposes. First, it validates the correctness of the overall pipeline including segmentation, normalization, model implementation, and training code, because even subtle bugs would typically manifest as non-trivial error rates. Second, it confirms that the HIT dataset contains sufficiently strong discriminative structure at the segment level under matched conditions. Therefore, the poor generalization observed later in cross-bearing experiments is unlikely to be caused solely by a fundamentally ill-posed classification problem; rather, it is closely associated with the mismatch between training and deployment domains.

Because the random split already yields saturated performance with a compact CNN, additional backbone comparisons under this protocol are not informative; we therefore focus stronger baselines on the cross-bearing setting where distribution shift is explicit.

4.3. Experiment 1: Three-Class Random Split

We next consider a more fine-grained three-class classification problem distinguishing normal, inner-race fault, and outer-race fault. The same random segment-level split strategy as in Experiment 0 is employed across all five bearings, so that the training and test segments are drawn from a matched segment-level distribution rather than from held-out bearing domains.

As shown in Table 4, under this setting, HitCNN1D maintains excellent performance, with the test accuracy again reaching 100%, and the confusion matrix remains perfectly diagonal, with 191/191 normal segments, 202/202 inner-race segments, and 90/90 outer-race segments correctly classified. All per-class F1-scores are equal to 1.0. These results

indicate that, under matched conditions, the proposed architecture is capable not only of detecting the presence of a fault but also of distinguishing between inner- and outer-race damage with negligible error.

Table 4. Experiment 1—random three-class classification.

Class	Precision	Recall	F1-Score	Support
Normal	1.0000	1.0000	1.0000	191
Inner	1.0000	1.0000	1.0000	202
Outer	1.0000	1.0000	1.0000	90
Overall accuracy	1.0000 (483/483)			

From a diagnostic viewpoint, Experiment 1 suggests that the multi-sensor features extracted by the CNN effectively encode both global energy patterns and more subtle spectral-temporal signatures associated with different fault locations. However, at the same time, the almost saturating metrics highlight a limitation of purely random splits that they drastically underestimate the difficulty of cross-machine deployment, and therefore cannot be the sole basis for evaluating algorithms intended for whole-engine health monitoring.

4.4. Experiment 2a: Cross-Bearing Binary Split (Normal vs. Inner)

To better approximate deployment on unseen hardware, Experiment 2a introduces a binary cross-bearing evaluation protocol that explicitly enforces distribution shift between the source and target domains. In the representative split, data1 normal and data3 inner-race fault samples are used as the source domain for training and validation, while data2 normal and data4 inner-race fault samples are held out as the target domain for final testing. This setup yields 1008 source segments, including 504 normal and 504 inner-race segments, and 954 target segments, including 450 normal and 504 inner-race segments. The target-domain bearing data are not used for training, validation, normalization-statistic estimation, or model selection.

Table 5 reports a representative single-run classification result under this cross-bearing protocol. The purpose of this table is not to claim statistical stability from a single seed, but to expose the dominant error mode in a transparent class-wise form. To assess seed-level variability, the same source–target split is repeated with five random seeds and summarized later. The multi-seed results further show that cross-bearing performance is not only low but also unstable, indicating strong sensitivity to optimization stochasticity under source–target mismatch.

For the baseline HitCNN1D, the representative result shows a near-complete collapse on both classes, with an overall accuracy of 0.0042. This indicates that the decision rule learned from the source bearings does not transfer reliably to the held-out target bearings in this representative run. More importantly, the normal class is not recognized at all in this run. In an on-line PHM system, such behavior would produce continuous or near-continuous false alarms when the target bearing is actually healthy.

The SE-enhanced model improves the representative accuracy to 0.1509 and recovers a small fraction of both normal and fault samples. This suggests that channel-wise recalibration can provide some structured benefit by reweighting intermediate feature channels. However, the absolute performance remains far below deployable levels, and the normal-class recall is still only 0.0711 in the representative run. Therefore, SE attention should be interpreted as a limited architectural enhancement rather than as a solution to cross-bearing domain shift.

The ResNet1D baseline achieves the highest representative accuracy among the three models in Table 5, reaching 0.2338, mainly because it recovers more inner-race fault samples.

Nevertheless, its normal-class recall remains 0.0000 in this run, corresponding to a false-alarm rate of 1.0000. This indicates that a deeper residual backbone can improve part of the fault sensitivity but still fails to identify healthy operation in the unseen target bearing. Thus, the result does not support deployment-ready generalization.

Table 5. Experiment 2a: cross-bearing binary classification for one representative run. The task is normal vs. inner-race fault detection.

(a) Baseline HitCNN1D				
Class	Precision	Recall	F1-Score	Support
Normal	0.0000	0.0000	0.0000	450
Inner	0.0083	0.0079	0.0084	504
Overall accuracy	0.0042 (4/954)			
Macro avg	0.0042	0.0040	0.0042	954
(b) HitCNN1D with SE attention				
Class	Precision	Recall	F1-Score	Support
Normal	0.0755	0.0711	0.0732	450
Inner	0.2113	0.2222	0.2166	504
Overall accuracy	0.1509 (144/954)			
Macro avg	0.1434	0.1467	0.1449	954
(c) ResNet1D				
Class	Precision	Recall	F1-Score	Support
Normal	0.0000	0.0000	0.0000	450
Inner	0.3314	0.4425	0.3789	504
Overall accuracy	0.2338 (223/954)			
Macro avg	0.1657	0.2212	0.1895	954

4.5. Experiment 2b: Extended Label-Space Cross-Bearing Evaluation

Experiment 2b extends the source-domain label space from binary detection to three output classes: normal, inner-race fault, and outer-race fault. In this setting, the source domain consists of data1 normal, data3 inner-race fault, and data5 outer-race fault samples, while the target domain consists of data2 normal and data4 inner-race fault samples. Therefore, this experiment should be interpreted as an extended label-space cross-bearing evaluation rather than a complete three-class target-domain fault-localization benchmark, because no outer-race target bearing is available in the held-out target set.

This design is still useful for evaluating deployment robustness. The model is trained to distinguish three health-state outputs in the source domain, but it must generalize to unseen target bearings whose samples belong only to the normal and inner-race classes. Any prediction of the outer-race class for target-domain normal or inner-race samples is counted as an error. Thus, Experiment 2b tests whether increasing label granularity in the source domain further destabilizes target-domain recognition.

Table 6 shows that both source-only models remain far from deployable performance under the extended label-space setting. The baseline HitCNN1D achieves only 0.0545 accuracy and 0.0536 macro F1-score, indicating severe target-domain recognition failure. The SE-enhanced model improves the accuracy to 0.2096 and the macro F1-score to 0.1803, but the absolute performance remains low. Therefore, the result should not be interpreted as successful three-class cross-bearing localization. Instead, it reinforces the main finding

that source-trained models can fail sharply when evaluated on unseen bearing domains, especially when the classifier is trained with a larger label space.

Table 6. Experiment 2b: extended label-space cross-bearing evaluation. The source domain includes normal, inner-race, and outer-race classes, while the held-out target domain contains normal and inner-race samples only.

Model	Test Accuracy	Macro F1-Score
HitCNN1D	0.0545	0.0536
HitCNN1D with SE attention	0.2096	0.1803

4.6. Ablation Study: Effect of RMS Normalization and Architecture

To isolate the effects of channel attention and RMS energy normalization under severe domain shift, we conduct an ablation study within the lightweight CNN family and evaluate four variants: (1) HitCNN1D, (2) HitCNN1D with RMS normalization, (3) HitCNN1D_SE, and (4) HitCNN1D_SE with RMS normalization. In addition, we report a stronger ResNet1D backbone as an extra baseline under the same cross-bearing protocol, while keeping RMS ablation within the CNN family to avoid unnecessary experimental expansion. Since cross-bearing results are sensitive to initialization and optimization stochasticity, we repeat each setting with five random seeds and report mean $\pm$ standard deviation. Across seeds, we keep the held-out target bearing fixed so that the reported variance reflects training randomness rather than a change of test domain.

Table 7 yields four engineering-relevant observations for the original cross-bearing split. First, the baseline CNN exhibits large seed-to-seed variance, indicating that cross-bearing generalization is not only weak but also unstable. This instability is consistent with severe source–target mismatch, where the optimizer may converge to qualitatively different decision rules depending on initialization.

Table 7. Multi-seed comparison under cross-bearing fault diagnosis: ablation on SE attention and RMS normalization within the lightweight CNN family, plus a stronger ResNet1D backbone baseline.

Model	Accuracy	Macro F1	Normal Recall
CNN (baseline)	0.1566 $\pm$ 0.1679	0.1251 $\pm$ 0.1295	0.0116 $\pm$ 0.0159
CNN + RMS	0.0711 $\pm$ 0.0850	0.0618 $\pm$ 0.0713	0.0000 $\pm$ 0.0000
SE-CNN	0.1692 $\pm$ 0.1288	0.1364 $\pm$ 0.0943	0.0124 $\pm$ 0.0278
SE-CNN + RMS	0.0839 $\pm$ 0.0863	0.0730 $\pm$ 0.0695	0.0000 $\pm$ 0.0000
ResNet1D	0.2218 $\pm$ 0.0955	0.1909 $\pm$ 0.0767	0.1169 $\pm$ 0.1549

Second, SE attention provides a modest improvement in this particular split, with slightly higher mean accuracy and macro F1-score than the plain CNN. However, the normal-class recall remains close to zero. Therefore, the gain from channel attention should be interpreted as limited rather than as a complete solution to cross-bearing domain shift. In deployment terms, the dominant failure mode is still the inability to recognize normal conditions in the unseen target bearing, which would translate into frequent nuisance alarms.

Third, per-segment RMS normalization consistently degrades the cross-bearing results and collapses normal-class recall to zero in the RMS-normalized variants. This indicates that aggressive energy normalization can suppress amplitude contrasts that remain informative for identifying normal operation under domain shift. From a deployment perspective, the RMS-normalized variants would label nearly every normal target-domain window as fault, which is unacceptable for on-line condition monitoring.

Fourth, the deeper ResNet1D baseline improves the average accuracy and macro F1-score relative to the lightweight CNN-family models in this split, but its deployment-oriented metrics remain poor. Over five random seeds, ResNet1D achieves an accuracy of 0.2218 ± 0.0955 and a macro F1-score of 0.1909 ± 0.0767 , while the normal-class recall remains low at 0.1169 ± 0.1549 . Given Equation (8), this corresponds to a high false-alarm rate of 0.8831 ± 0.1549 . Therefore, within the tested source-only CNN and ResNet1D backbones, increasing backbone capacity alone does not eliminate the cross-bearing degradation under the present protocol.

4.7. Robustness Across Multiple Cross-Bearing Permutations

The original Experiment 2a uses one source–target bearing split. To examine whether the observed degradation is an artifact of this particular split, we further conduct an additional permutation study for the binary normal-vs.-inner-race task. Four cross-bearing permutations are evaluated: source data1+data3 to target data2+data4, source data1+data4 to target data2+data3, source data2+data3 to target data1+data4, and source data2+data4 to target data1+data3. For each permutation, the source domain is split into training and validation subsets with a stratified 80%/20% ratio, while the target domain is used only for final testing. Each setting is repeated with five random seeds. The permutation study evaluates the source-only CNN, SE-CNN, and ResNet1D backbones without RMS normalization.

Table 8 further indicates that the cross-bearing degradation is not solely attributable to a single unfavorable source–target split. Across all four permutations, the source-only CNN and SE-CNN achieve similar aggregate performance, with accuracy around 0.36 and normal-class recall below 0.06. This indicates that the benefit of SE attention is split-dependent and limited under the present source-only training protocol. ResNet1D provides a modest average improvement, increasing the aggregate accuracy to 0.3905 and the macro F1-score to 0.3197. However, its normal-class recall remains only 0.1197 ± 0.1816 , corresponding to a false-alarm rate of 0.8803 ± 0.1816 . Therefore, even the stronger residual backbone remains far from deployment-ready cross-bearing generalization.

Table 8. Aggregate performance over four binary cross-bearing permutations and five random seeds per permutation. Values are reported as mean $\pm$ standard deviation over 20 runs for each model.

Model	Accuracy	Macro F1	Balanced Acc.	Normal Recall	FAR	Miss Rate
CNN	0.3593 ± 0.1680	0.2815 ± 0.1362	0.3510 ± 0.1635	0.0517 ± 0.0925	0.9483 ± 0.0925	0.3496 ± 0.2932
SE-CNN	0.3592 ± 0.1592	0.2819 ± 0.1244	0.3509 ± 0.1549	0.0566 ± 0.1087	0.9434 ± 0.1087	0.3548 ± 0.2865
ResNet1D	0.3905 ± 0.1460	0.3197 ± 0.1364	0.3832 ± 0.1442	0.1197 ± 0.1816	0.8803 ± 0.1816	0.3533 ± 0.2712

These additional results strengthen the central finding of this study. Random segment-level splits produce near-ceiling performance because the training and test samples share bearing-specific characteristics. In contrast, when the model is evaluated on unseen bearings, the learned source-domain decision boundary does not transfer reliably. The dominant practical consequence is not merely a reduction in accuracy, but a persistent tendency to classify target-domain normal operation as fault. This behavior would generate frequent false alarms in an on-line PHM system and highlights the need for future work on domain adaptation, domain generalization, calibration, or sequential decision logic.

5. Discussion and Deployment Considerations

5.1. Lessons from Random and Cross-Bearing Evaluation

Experiments 0 and 1 show near-ceiling performance under random segment-level splits, where both the binary detection task and the three-class discrimination task reach

100% test accuracy. These results indicate that the HIT inter-shaft bearing benchmark contains highly discriminative information at the segment level when training and test data are drawn from the same bearing domain. They also validate the basic correctness of the implemented pipeline, including segmentation, normalization, model training, and inference. However, such saturated performance should not be interpreted as evidence of deployment-ready generalization. Because windows from the same physical bearing can appear in both training and test subsets under random splitting, the evaluation may allow bearing-specific or run-specific signatures to be shared between the two subsets. Therefore, random segment-level testing mainly reflects in-domain interpolation and can lead to overly optimistic conclusions about real-world performance.

The contrast becomes clear under the cross-bearing protocol of Experiment 2a. When the target bearing domain is held out, the lightweight CNN-family models exhibit low accuracy, low macro F1-score, and large seed-to-seed variability. In particular, the baseline HitCNN1D yields low mean accuracy and almost vanishing normal-class recall under the representative cross-bearing split. This failure mode is asymmetric. The dominant error is not merely a uniform reduction of classification quality, but a tendency to classify target-domain normal windows as fault. In an on-line PHM system, this behavior would lead to frequent nuisance alarms and could reduce operator trust in the diagnostic system.

SE attention provides a limited architectural improvement in the representative cross-bearing split, but its effect is not sufficient to remove the deployment risk. The SE-enhanced model slightly improves mean accuracy and macro F1-score in the original split, yet the normal-class recall remains close to zero. The additional permutation study further shows that the aggregate performance of CNN and SE-CNN is similar across four source–target combinations. Therefore, SE attention should be interpreted as a lightweight channel-recalibration mechanism with split-dependent benefits, rather than as a domain-robustness solution.

The RMS ablation also highlights that preprocessing is not a neutral implementation detail. Per-segment RMS normalization is often used to reduce amplitude variability, but in this benchmark it degrades the cross-bearing results and collapses normal-class recall in the RMS-normalized variants. One possible explanation is that amplitude-related information remains useful for distinguishing normal and fault states under the HIT setting, and aggressive per-window energy normalization may suppress this information. This observation suggests that normalization strategies should be evaluated as part of the deployment protocol rather than treated as purely numerical preprocessing.

A natural question is whether the poor cross-bearing performance mainly reflects limited model capacity. To examine this issue, we include a deeper ResNet1D baseline under the same source-only split, preprocessing, and model-selection protocol. ResNet1D improves the average accuracy and macro F1-score relative to the lightweight CNN-family models in the representative split, and it also provides the best aggregate accuracy in the permutation study. However, its deployment-oriented indicators remain unfavorable. In the representative split, its normal-class recall remains low and the false-alarm rate remains high. Across the four permutation splits, its aggregate false-alarm rate is also still far from acceptable for deployment. These outcomes indicate that, within the tested source-only CNN and ResNet1D backbones, increasing backbone capacity alone does not eliminate the cross-bearing degradation. The dominant bottleneck under the present protocol appears to be source–target distribution mismatch, which is consistent with the dataset-shift and domain-adaptation perspectives.

5.2. Implications for Deployment: False Alarms, Misses, and Calibration

The cross-bearing findings have direct implications for practical PHM deployment. In a realistic monitoring system, the diagnostic input is a continuously updated multi-sensor time-domain window, and the output is a per-window fault probability followed by an alarm or diagnostic decision. Under this workflow, overall accuracy alone is insufficient to characterize operational risk. A model can obtain non-trivial accuracy while still generating frequent nuisance alarms if it fails to recognize normal operation in an unseen target domain. This is why balanced accuracy, false-alarm rate, miss rate, and class-wise recall are necessary complements to overall accuracy.

False alarms and missed detections represent different but coupled engineering risks. A high false-alarm rate threatens operability by triggering nuisance alerts, alarm fatigue, unnecessary inspection, and reduced trust in the monitoring system. A high miss rate threatens safety by allowing true fault signatures to remain undetected. The cross-bearing results show that random-split-validated models can exhibit severe false-alarm behavior when applied to unseen bearings, even when the same diagnostic labels are used. Therefore, a single-window classifier trained only on source-domain bearings should not be directly treated as a deployment-ready alarm system.

A deployable PHM system would require additional decision layers beyond the source-only classifier evaluated in this paper. Possible mechanisms include threshold calibration under explicit false-alarm constraints, temporal smoothing of window-level predictions, sequential decision logic, uncertainty estimation, and domain-robust representation learning. Such mechanisms are outside the scope of the present baseline study, but the proposed evaluation protocol and reported metrics provide a concrete basis for testing them. In particular, future methods should be evaluated not only by mean accuracy, but also by their ability to reduce false alarms while preserving fault sensitivity on held-out target bearings.

5.3. Relation to Domain Adaptation and Domain Generalization

The present work does not propose a domain-adaptation or domain-generalization algorithm. Instead, it establishes a source-only baseline and a deployment-oriented evaluation protocol for aero-engine-relevant multi-sensor vibration diagnosis. This distinction is important. The poor cross-bearing results should not be read as evidence that deep learning is unsuitable for aero-engine PHM. Rather, they indicate that conventional source-only training and random segment-level validation are insufficient for assessing deployability under unseen-bearing conditions.

The results also clarify how future domain-robust methods should be evaluated. Domain-adversarial learning, discrepancy minimization, self-supervised pretraining, contrastive learning, physics-aware representation learning, and target-domain calibration are all plausible directions for improving robustness under source–target mismatch. However, these methods should be tested under bearing-level or engine-level held-out protocols, and their gains should be reported using deployment-oriented metrics. Otherwise, improvements under random splits may still reflect in-domain interpolation rather than genuine cross-domain transfer.

From this perspective, the contribution of the present study is not a new adaptation mechanism, but a reproducible diagnostic baseline and a protocol-level warning. Random segment-level splits can make the benchmark appear solved, whereas cross-bearing evaluation reveals substantial false-alarm risk. This finding provides a stricter reference point for subsequent work that aims to improve cross-bearing or cross-engine generalization.

5.4. Limitations and Future Work

This study is intentionally scoped as a protocol-driven baseline on a rare aero-engine-relevant multi-sensor vibration benchmark. Several limitations should be stated clearly. First, the experiments are conducted on one public HIT inter-shaft bearing dataset and a limited set of bearing fault classes. Therefore, the conclusions should be interpreted as evidence about cross-bearing shift on this benchmark rather than universal claims about all aero-engine vibration diagnosis tasks. Second, the evaluated models are trained under standard source-only empirical risk minimization without explicit domain adaptation, physics-guided constraints, target-domain calibration, or extensive data augmentation. Third, although this revision adds multiple binary cross-bearing permutations and reports multi-seed statistics, broader validation across additional aero-engine rigs, operating regimes, and fault categories remains necessary.

These limitations motivate several concrete directions. One direction is to incorporate domain robustness mechanisms, including domain-adversarial objectives, discrepancy minimization, self-supervised pretraining, and other transfer learning strategies that explicitly target source-to-target mismatch. A second direction is to integrate operating-condition information and physics-aware priors so that learned representations are encouraged to align with fault-relevant dynamics rather than incidental bearing identity. A third direction is to develop deployment-oriented decision logic, including calibrated operating thresholds, temporal aggregation, uncertainty estimation, and alarm persistence rules. Finally, future work should validate such methods on additional aero-engine or whole-engine datasets and report performance under explicit false-alarm constraints to quantify robustness across a broader range of target domains.

6. Conclusions

This study revises the evaluation of multi-sensor aero-engine-relevant vibration diagnosis from a random segment-level testing problem to a cross-bearing generalization problem. Using the HIT inter-shaft bearing benchmark, we show that random splits can yield near-ceiling performance, whereas bearing-level held-out evaluation reveals severe degradation and high false-alarm risk. The additional permutation study further indicates that this degradation is not solely caused by one unfavorable source–target split. Within the tested source-only CNN, SE-CNN, and ResNet1D backbones, increasing backbone capacity and adding channel attention provide limited or split-dependent benefits but do not eliminate the cross-bearing false-alarm problem. These results establish a reproducible source-only baseline and motivate future work on domain adaptation, domain generalization, calibration, and sequential decision logic for deployment-ready aero-engine PHM.

Author Contributions: Conceptualization, Y.X. and C.H.; methodology, Y.X. and J.S.; software, Y.X. and H.P.; validation, Y.X. and C.W.; formal analysis, Y.X.; investigation, Y.X.; resources, C.H. and J.Z.; data curation, C.H. and J.Z.; writing—original draft preparation, Y.X.; visualization, Y.X.; supervision, C.H. and J.Z.; project administration, C.H. and J.Z.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research on the Construction of Airborne of the Chinese Academy of Sciences from Jiaxian Sun grant number 1Z2024000132.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiments in this study use the public HIT inter-shaft bearing dataset, which is available at <https://github.com/HouLeiHIT/HIT-dataset> (accessed on 15 September 2025). No self-acquired dataset is used in the reported experiments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. ISO 13373-1:2002; Condition Monitoring and Diagnostics of Machines—Vibration Condition Monitoring—Part 1: General Procedures. International Standard. International Organization for Standardization (ISO): Geneva, Switzerland, 2002.
2. ISO 13373-2:2016; Condition Monitoring and Diagnostics of Machines—Vibration Condition Monitoring—Part 2: Processing, Analysis and Presentation of Vibration Data; International Standard; International Organization for Standardization (ISO): Geneva, Switzerland, 2016.
3. Matthaiou, I.; Khandelwal, B.; Antoniadou, I. Vibration monitoring of gas turbine engines: Machine-learning approaches and their challenges. *Front. Built Environ.* **2017**, *3*, 54.
4. ISO 10816-4:2009; Mechanical Vibration—Evaluation of Machine Vibration by Measurements on Non-Rotating Parts—Part 4: Gas Turbine Sets with Fluid-Film Bearings. International Organization for Standardization (ISO): Geneva, Switzerland, 2009.
5. Clifton, D.A. *Condition Monitoring of Gas-Turbine Engines*; Transfer Report; Department of Engineering Science, University of Oxford: Oxford, UK, 2006.
6. Antoni, J. The spectral kurtosis: A useful tool for characterising non-stationary signals. *Mech. Syst. Signal Process.* **2006**, *20*, 282–307.
7. Muszyńska, A. Rotor-to-stationary element rub-related vibration phenomena in rotating machinery—Literature survey. *Shock Vib. Dig.* **1989**, *21*, 3–11.
8. Goldman, P.; Muszyńska, A. Rotor-to-stator, rub-related, thermal/mechanical effects in rotating machinery. *Chaos Solitons Fractals* **1995**, *5*, 1579–1601.
9. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520.
10. Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510.
11. Liu, R.; Yang, B.; Zio, E.; Chen, X. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process.* **2018**, *108*, 33–47.
12. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587.
13. Ince, T.; Kiranyaz, S.; Eren, L.; Askar, M.; Gabbouj, M. Real-time motor fault detection by 1-D convolutional neural networks. *IEEE Trans. Ind. Electron.* **2016**, *63*, 7067–7075.
14. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2015; pp. 448–456.
15. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
16. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York City, NY, USA, 2018; pp. 7132–7141.
17. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Aroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929.
18. Quionero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N.D. *Dataset Shift in Machine Learning*; MIT Press: Cambridge, MA, USA, 2009.
19. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
20. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
21. Li, W.; Huang, R.; Li, J.; Liao, Y.; Chen, Z.; He, G.; Yan, R.; Gryllias, K. A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges. *Mech. Syst. Signal Process.* **2022**, *167*, 108487.
22. Tang, S.; Ma, J.; Yan, Z.; Zhu, Y.; Khoo, B.C. Deep transfer learning strategy in intelligent fault diagnosis of rotating machinery. *Eng. Appl. Artif. Intell.* **2024**, *134*, 108678.
23. NASA. CMAPSS Jet Engine Simulated Data (C-MAPSS): Turbofan Engine Degradation Simulation Data Set. Available online: <https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data> (accessed on 5 February 2026).
24. Arias Chao, M.; Kulkarni, C.; Goebel, K.; Fink, O. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data* **2021**, *6*, 5.
25. Case Western Reserve University. Bearing Data Center. Available online: <https://csegroups.case.edu/bearingdatacenter> (accessed on 5 February 2026).
26. Hou, L.; Yi, H.; Jin, Y.; Gui, M.; Sui, L.; Zhang, J.; Chen, Y. Inter-shaft bearing fault diagnosis based on aero-engine system: A benchmarking dataset study. *J. Dyn. Monit. Diagn.* **2023**, *2*, 228–242.

27. Wang, B.; Lei, Y.; Li, N.; Li, N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Trans. Reliab.* **2020**, *69*, 401–412.
28. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York City, NY, USA, 2016; pp. 770–778.
30. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates Inc.: Red Hook, NY, USA, 2019. Available online: <https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library> (accessed on 12 May 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.